Imputation, Sampling, Allocation in Census 2000

State Data Center Meeting
September 22, 2004

Outline:

- Basic Survey Concepts
- 100 % Data Imputations
- Census Sampling & the Long Form
- Long Form Estimation
- Missing Sample Data
- Census Confidence Intervals
- Conclusion

Outline:

- **Basic Survey Concepts**
- 100 % Data Imputations
- Census Sampling & the Long Form
- Long Form Estimation
- Missing Sample Data
- Census Confidence Intervals
- Conclusion

Some Definitions

- Observation Unit or Element
- Sampling unit
- Sampling frame

Observation Unit or Element

The object on which a measurement is to be taken

- – Individual
- – Household
- – Housing Unit

Sampling Unit

The unit we actually select in the sample

IDs on the Decennial Master Address List

Probability Sampling

Sample selection where every sampling unit has a known probability of selection

Stratification

Divide frame into groups with similar units

- – Each group is known as a stratum

Domains

Often you will want to make separate estimates for sub-groups in the population; these are known as domains.

Stratification often can help support separate estimates for domains.

Simple Estimation

Generally:

Multiply each sample unit's value by the number of cases it "represents."

Simple Estimation Example

- Sample size is 64
- There are 128,000 sampling units on the frame
- Each sample unit "represents" 2,000 units including itself

Simple Estimation Example

If the total value for the sample is:

$8,192

Then the total for the population is:

2000 x 8,192

Or

$16,384,000

Ratio Estimation

Sometimes we have outside (auxiliary) information on all units in the target population or frame

Such as:

Results from Earlier Census
Administrative Records

13

Simple Ratio Example

Total current value for the sample is
$8,192

Value for <u>sample units</u> in last census was
$7,314

Value in our sample grew by 12%,
8,192/7,314 = 1.12

14

Ratio Estimation Example

- Estimated growth 12% or 1.12 to 1
- Total Target Population value at time of last census:
  15,473,900
- So our estimate would be
  15,473,900 * 1.12 = 17,330,768

15

Sampling Errors

Since we selected only one subset of the target population, our results would be different if we had selected another of the many possible subsets

Sampling Errors can be measured

16

Sampling Errors

Sampling errors can be reported in terms of a "confidence interval"

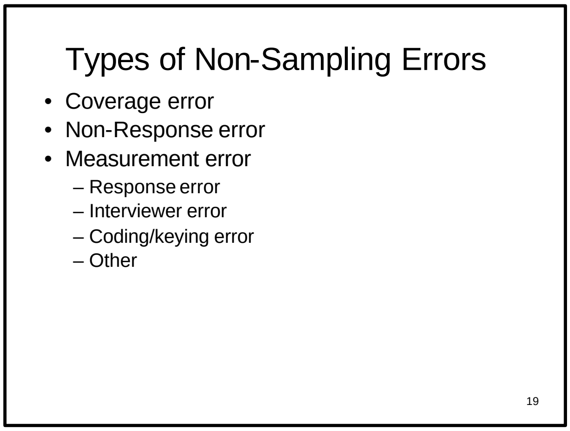
The confidence interval is the smallest range around the estimate that is likely to include the true value with a specified probability.

17

Two Examples of Confidence intervals

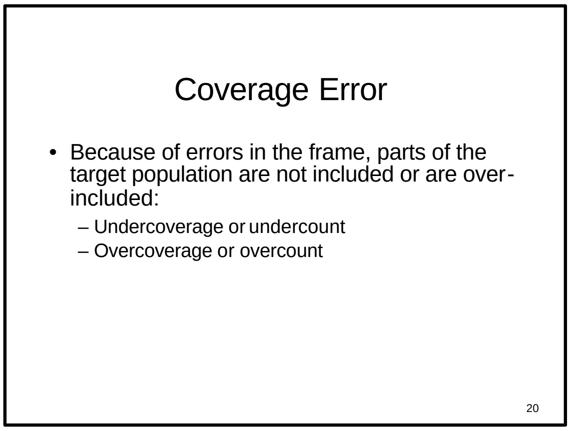
- Proportion employed is 34% ± 4%
  or from 30 to 38%
- Total housing value is 16.4 ± 0.7 million
  or from 15.7 to 17.1 million

18



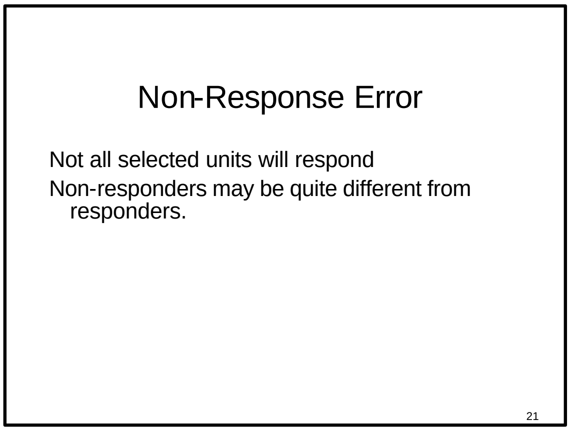
Types of Non-Sampling Errors

- Coverage error
- Non-Response error
- Measurement error
  - Response error
  - Interviewer error
  - Coding/keying error
  - Other



Coverage Error

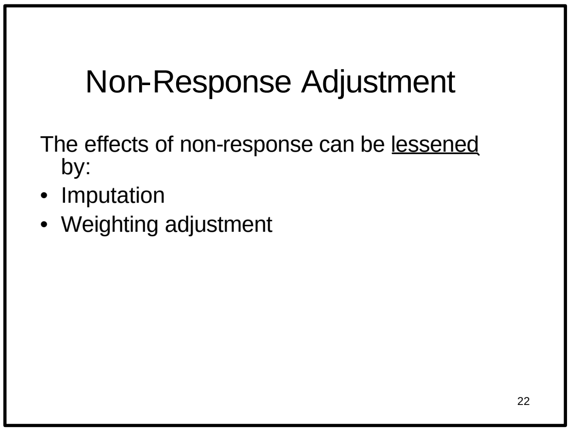
- Because of errors in the frame, parts of the target population are not included or are over-included:
  - Undercoverage or undercount
  - Overcoverage or overcount



Non-Response Error

Not all selected units will respond

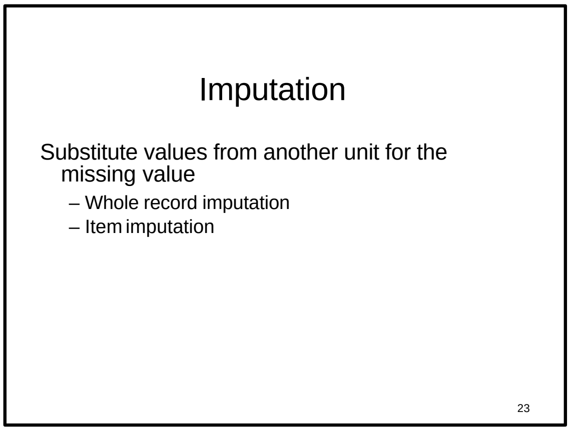
Non-responders may be quite different from responders.



Non-Response Adjustment

The effects of non-response can be lessened by:

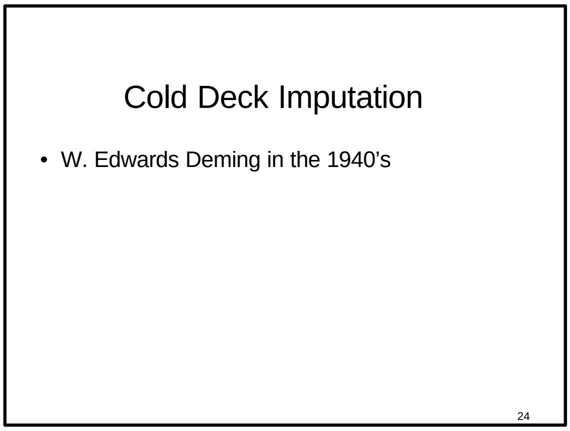
- Imputation
- Weighting adjustment



Imputation

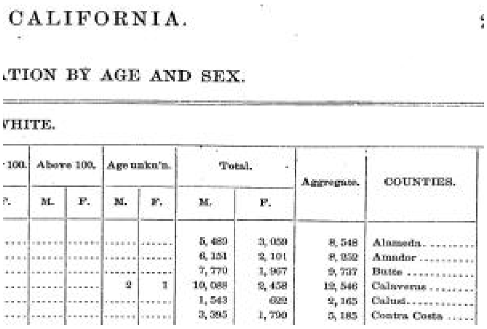
Substitute values from another unit for the missing value

- Whole record imputation
- Item imputation



Cold Deck Imputation

- W. Edwards Deming in the 1940’s

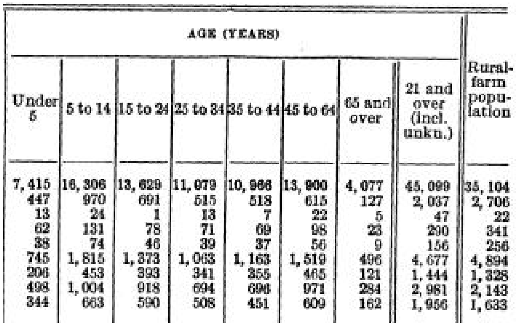


CALIFORNIA.

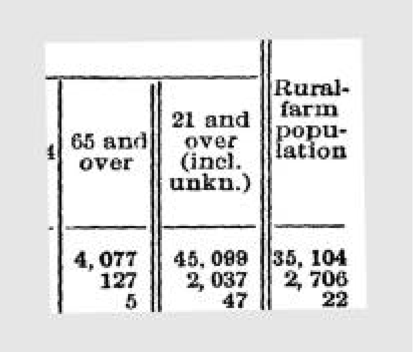
.TION BY AGE AND SEX.

WHITE.

100.	Above 100.		Age unkn'n.		Total.		Aggregate.	COUNTIES.
F.	M.	F.	M.	F.	M.	F.		
					5, 489	3, 059	8, 548	Alameda
					6, 151	2, 101	8, 252	Amador
					7, 770	1, 967	9, 737	Butte
			2	1	10, 088	2, 458	12, 546	Calaveras
					1, 543	622	2, 165	Colusa
					3, 395	1, 790	5, 185	Contra Costa



AGE (YEARS)								
Under 5	5 to 14	15 to 24	25 to 34	35 to 44	45 to 64	65 and over	21 and over (incl. unkn.)	Rural-farm popu-lation
7, 415	16, 306	13, 629	11, 079	10, 966	13, 900	4, 077	45, 099	35, 104
447	970	691	515	518	615	127	2, 037	2, 706
13	24	1	13	7	22	5	47	22
62	131	78	71	69	98	23	290	341
38	74	46	39	37	56	9	156	256
745	1, 815	1, 373	1, 063	1, 163	1, 519	496	4, 677	4, 894
206	453	393	341	355	465	121	1, 444	1, 328
498	1, 004	918	694	696	971	284	2, 981	2, 143
344	663	590	508	451	609	162	1, 956	1, 633



65 and over	21 and over (incl. unkn.)	Rural-farm popu-lation
4, 077	45, 099	35, 104
127	2, 037	2, 706
5	47	22

Hot Deck Imputation

- Experience of 1960 Census
- Modified Hot Deck

Weight Adjustment

- Simple Weight Adjustment
- Post-Stratification
- Raking

Simple Weight Adjustment

Total population:	1000 households
Sampled:	250 households
Simple weight	4
Responded	200 households
Non-Response	
Adjusted Weight	5

Post-stratification

Total population:	1000 households
Sampled:	250 households
Responded	200 households 300 people 100 Men 200 Women

Post-stratification

Responded	200 households 300 people 100 Men 200 Women
Known True number	1000 Households 2200 people 1000 Men 1200 Women

Post-stratifed Adjusted Weights

Households	1000/200 = 5
Men	1000/100 = 10
Women	1200/200 = 6

Purpose of Post-stratification

- Decrease variance
- Decrease non-response bias

Choosing Post-strata

- Large enough to decrease variance
- Small enough to decrease non-response bias.

Raking

Raking 2

Raking 3

Outline:

- Basic Survey Concepts
- **100 % Data Imputations**
- Census Sampling & the Long Form
- Long Form Estimation
- Missing Sample Data
- Census Confidence Intervals
- Conclusion

Census Processing Basics

- Address List Development
- Mail Out/Mail Back
- Non-Response Follow-up

Missing Data in 1850

Some of the returns from California have not yet been received. Assuming population of California to be 165,000, (which we do partly by estimate,) the total number of inhabitants of the United States was….23,263,488.

Report of the Superintendent of the Census, December 1852

(page 129)

Four Kinds of Non-Response for 100% Data Items

- Missing Population Count
- Population Count Only
- Missing Person Information
- Missing Items

Missing Population Count

- Unresolved status
- Unresolved occupancy
- Unknown household size

Status Imputation

Impute status based on occupied, vacant, or delete enumerator forms controlled by single vs. multi-unit

- 415,892
- 0.15%

Occupancy Imputation

Impute status based on occupied or vacant enumerator forms controlled by single vs. multi-unit

- 260,652
- 0.09%

Household Size Imputation

Impute status based on occupied enumerator forms controlled by single vs. multi-unit

- 495,600
- 0.18%

Total Count Imputation

- 1,172,144
- 0.42%

Utah v. Evans

	Number	%
• Utah	5,395	0.24
• North Carolina	32,457	0.40

Percent of Apportionment Population

Pop Count Only

Households with known population count but no data defined persons

Data Defined Person Record

At least two Characteristics
- Relationship
- Sex
- Race
- Hispanic Origin
- Age or Date of Birth
- Name
  - 3 Characters in name fields

Pop Count Only

2,269,010 Person Records
0.81%

Substitutions

Households where all person records are substituted from another household

No person record in household can be data-defined

Term and measure applies only to households

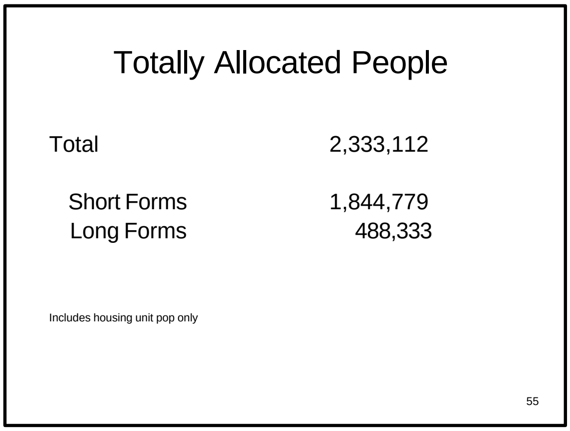
Whole Household Substitutions

	Number	%
Count Imputations	1,172,144	0.42%
Pop Count Only	2,269,010	0.81%
Total	3,441,154	1.22%

Some Census Jargon

- Substitutions: No data-defined people in household
- Totally allocated person record: Person record not data-defined, but others in household were
- Item Imputation: Record was data-defined but missing some values

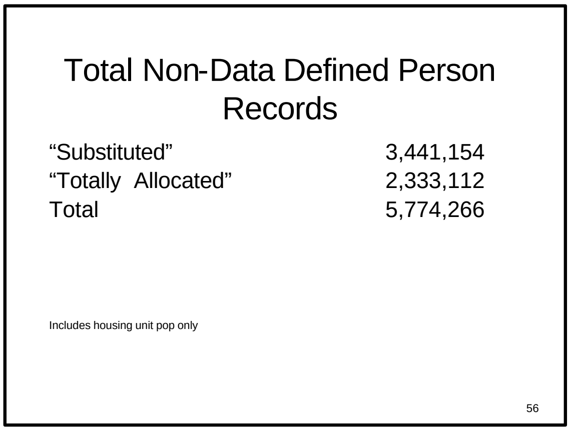
- NOT CONSISTENTLY USED



Totally Allocated People

Total	2,333,112
Short Forms	1,844,779
Long Forms	488,333

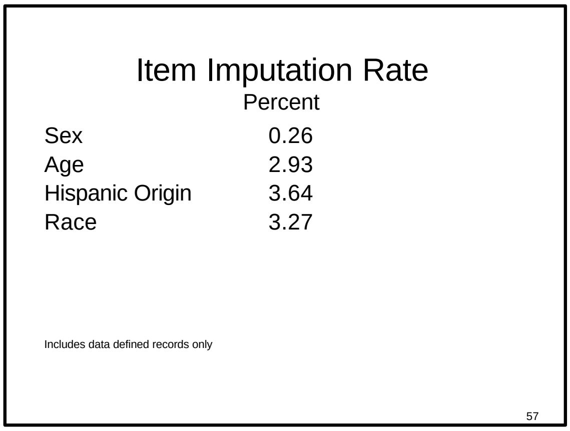
Includes housing unit pop only



Total Non-Data Defined Person Records

“Substituted”	3,441,154
“Totally Allocated”	2,333,112
Total	5,774,266

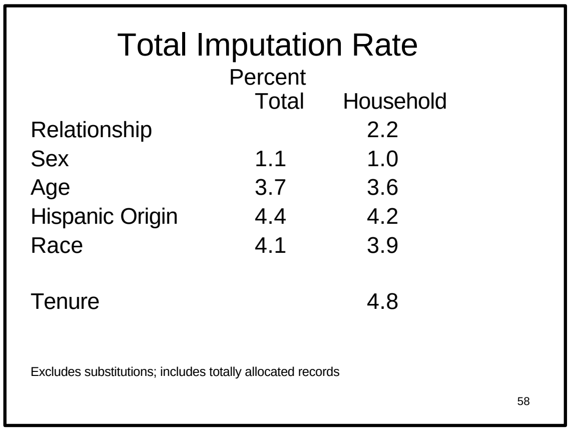
Includes housing unit pop only



Item Imputation Rate

	Percent
Sex	0.26
Age	2.93
Hispanic Origin	3.64
Race	3.27

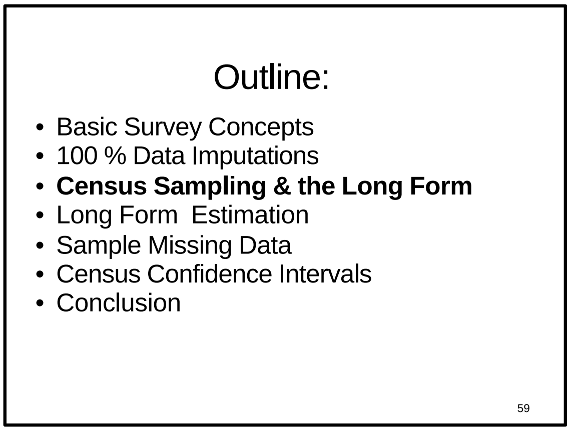
Includes data defined records only



Total Imputation Rate

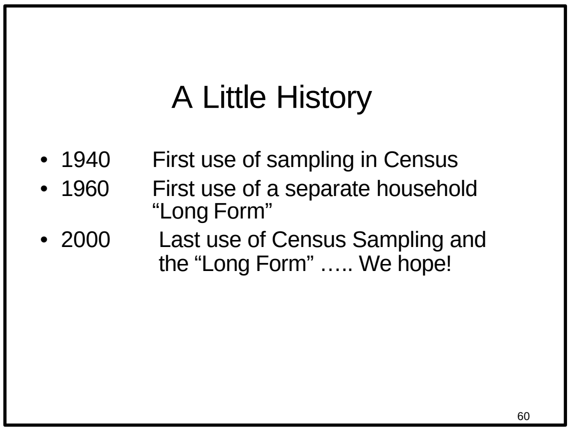
	Percent	
	Total	Household
Relationship		2.2
Sex	1.1	1.0
Age	3.7	3.6
Hispanic Origin	4.4	4.2
Race	4.1	3.9
Tenure		4.8

Excludes substitutions; includes totally allocated records



Outline:

- Basic Survey Concepts
- 100 % Data Imputations
- **Census Sampling & the Long Form**
- Long Form Estimation
- Sample Missing Data
- Census Confidence Intervals
- Conclusion



A Little History

1940	First use of sampling in Census
1960	First use of a separate household “Long Form”
2000	Last use of Census Sampling and the “Long Form” ….. We hope!

Long Form Sampling Entities

- Counties
- Cities
- Incorporated places
- School Districts
- American Indian reservations
- Certain other special cases

LFSE Sampling Rates

Housing Units	Sampling Rate
1 to 799	1 in 2
800 to 1199	1 in 4
1200 to 1999	1 in 6
Over 2000	1 in 8

Assignment of Blocks

If a block was in more than one LFSE, it got the sample weight of the sampling rate of the smallest LFSE.

Expected Sample Sizes

Housing Units	Sample Size
1 to 799	0 to 400
800 to 1199	200 to 300
1200 to 1999	200 to 334
Over 2000	250 or more

Realized Sample

- Different final HU count in area
- Long Form HU Non-response

Outline:

- Basic Survey Concepts
- 100 % Data Imputations
- Census Sampling & the Long Form
- **Long Form Estimation**
- Missing Sample Data
- Census Confidence Intervals
- Conclusion

Long Form Data Defined Person Record

- "100 % Data" Defined
  - Two characteristics possibly including "name"
- At least two nonempty sample data items

Long Form Data Defined Occupied Housing Unit Record

At least one long form data defined person record

Weighting Areas

- Formed, if possible, within Tabulation Block groups
- Cannot cross county boundary
- 200 Data Defined Long Forms
- 400 Sample Person Records

- 65,343 weighting areas were used

Initial Weight Adjustment

Calculate Observed Sampling Weight

Total People
DIVIDED BY
Sample Data Defined People

Similar process for Occupied and for Vacant HUs

Iterative Ratio Estimation

A series of ratio estimation procedures to force sample totals to agree with certain 100 percent data item totals.

Stage 1: Type of Household

- Families with own children under age 18
  - Household size 2, 3, 4, 5, 6-7, 8 plus
- Families without own children under 18
  - Household size 2, 3, 4, 5, 6-7, 8 plus
- All other housing units
  - Single
  - 2, 3, 4, 5, 6- 7, 8 plus
- People in Group Quarters
- Service Based Enumerations

Stage 2: Sampling Type

- 1 in 2
- 1 in 4
- 1 in 6 and 1 in 8

Stage 3: Householder Status

- Householder
- Nonhouseholder

Stage 4: Age/Sex/Race/Origin

- Age (13 Five-year groups)
    by
- Sex (2 groups)
    by
- Black, AIAN, Asian,NHPI, White, SOR (6)
    by
- Hispanic, Not of Hispanic origin   (2)

Collapsing

- Each category had to have at least 10 sample records and meet other minimum size requirements.
- Otherwise, it was combined with another group in the weighting area.
- Collapsing was a large contributor to sample vs. census differences

Iterative Ratio Estimation

At end of each stage, sample population estimates will exactly equal 100% Population counts for the collapsed groups in that stage

Iterative Ratio Estimation

- After Stage 1, estimates will exactly equal counts by type of household
- After Stage 2, estimates will exactly equal counts by Sampling type, but will no longer exactly equal counts by type of household
- After Stage 3, estimates will exactly equal counts by Householder status, but no longer exactly equal counts by Sampling Type
- Etc.

Iterative Ratio Estimation

Go through each stage again until the changes are very small.

Iterative Ratio Estimation

Separate weighting process for

- Person Records
- Occupied Housing Units
- Vacant Housing Units

Final Long Form Weight

- The weights are converted to Integers
- Each long form person will have a weight, i.e. represent so many other people
- Each long form householder will have a weight, i.e. represent so many other householders
- Each long form occupied housing unit will have a weight
- Each long form vacant housing unit will have a weight

Weighting Area Results

Estimates of total population, occupied and vacant housing units will agree exactly for areas of weighting area or larger

Estimates will agree exactly with collapsed counts for the last stage of fitting

(age, race, Hispanic origin, sex)

Estimates will be very close for the groups for the other weighting stages

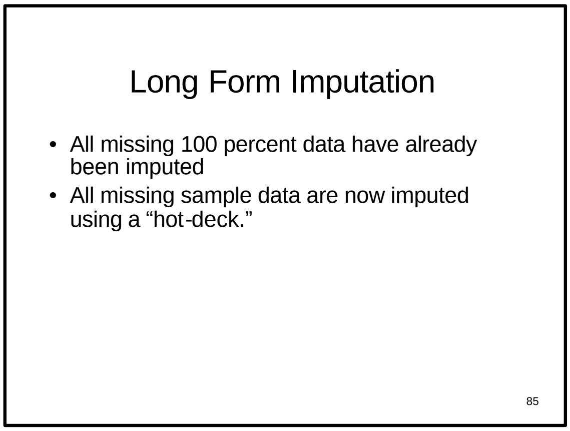
Results for other Areas

Estimates will agree exactly with collapsed counts for large areas made up of weighting areas

Estimates will differ from counts for areas smaller than a Weighting Area.

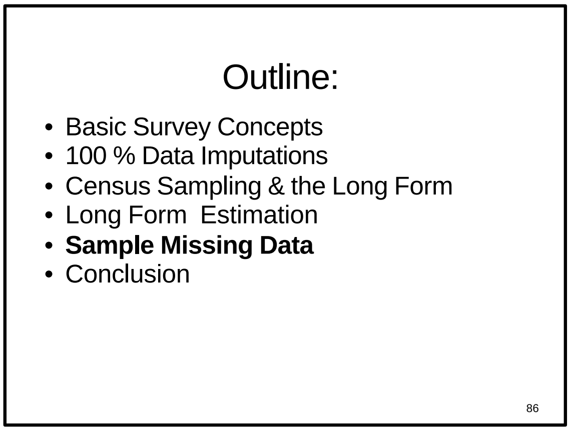
Household vs Occupied Housing Unit

- Household data receive the weight of the householder
- Data for occupied housing units receive the weight for the housing unit.
- The two may differ



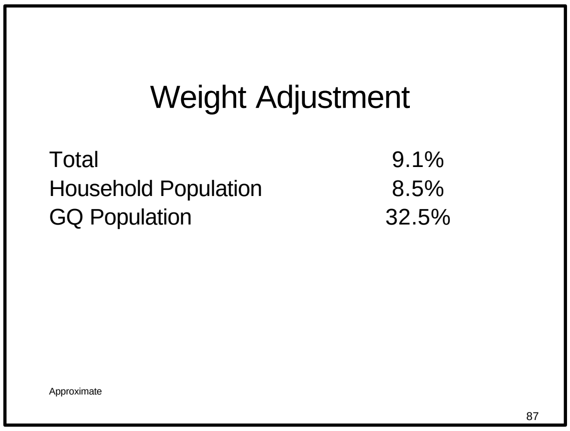
Long Form Imputation

- All missing 100 percent data have already been imputed
- All missing sample data are now imputed using a “hot-deck.”



Outline:

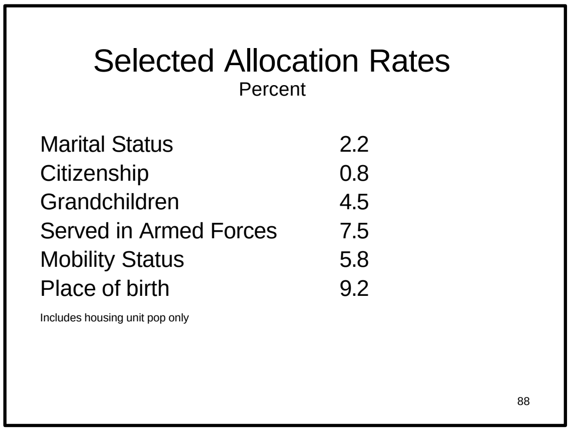
- Basic Survey Concepts
- 100 % Data Imputations
- Census Sampling & the Long Form
- Long Form Estimation
- **Sample Missing Data**
- Conclusion



Weight Adjustment

Total	9.1%
Household Population	8.5%
GQ Population	32.5%

Approximate

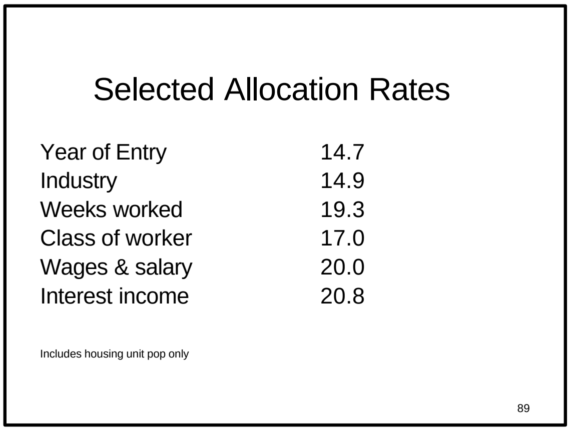


Selected Allocation Rates

Percent

Marital Status	2.2
Citizenship	0.8
Grandchildren	4.5
Served in Armed Forces	7.5
Mobility Status	5.8
Place of birth	9.2

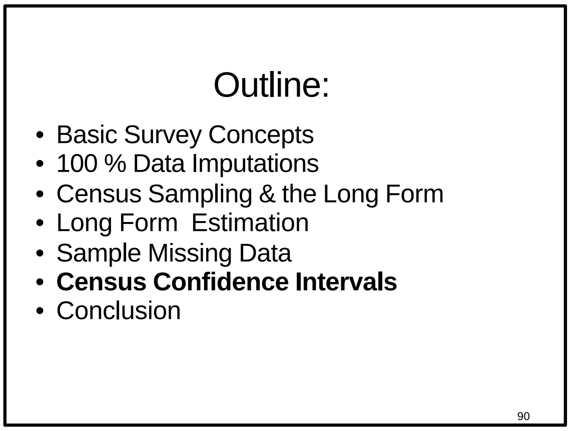
Includes housing unit pop only



Selected Allocation Rates

Year of Entry	14.7
Industry	14.9
Weeks worked	19.3
Class of worker	17.0
Wages & salary	20.0
Interest income	20.8

Includes housing unit pop only



Outline:

- Basic Survey Concepts
- 100 % Data Imputations
- Census Sampling & the Long Form
- Long Form Estimation
- Sample Missing Data
- **Census Confidence Intervals**
- Conclusion

Generalized Standard Errors

Estimated Totals

The Standard Error will depend upon three factors:

1. Size of population area
2. Estimated Total
3. Observed Sampling Rate

Generalized Standard Errors

Estimated Totals

Step 1

Look up unadjusted sampling error using Table A from Chapter 8: Accuracy of Data.

Generalized Standard Errors

Estimated Totals

Step 1

Oxnard Population

Total	170,358
Less than 5th grade	10,685

From Table A:

Unadjusted Standard Error	219

Generalized Standard Errors

Estimated Totals

Step 2

Obtain observed sampling rate from Table P4 or H4

Total Pop (100% count)	170,358
Unweighted count	22,706

Sampling Rate

22,706 / 170,358 = 13.3

Generalized Standard Errors

Estimated Totals

Step 3

Obtain Standard Error Design Factors from Table C

Row:	Educational Attainment Row
Column:	Less than 15% (13.3)

Factor = 1.3

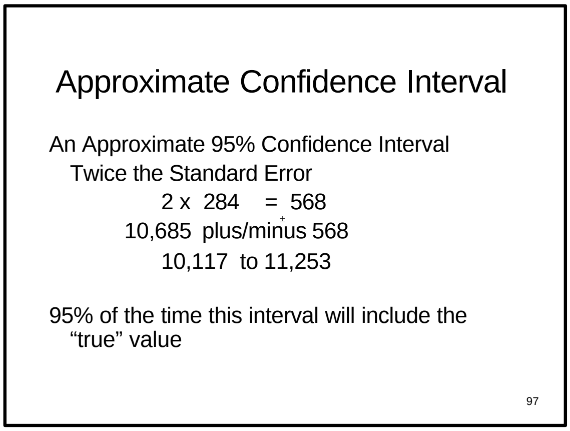
Generalized Standard Errors

Estimated Totals

Step 4

Compute the approximate Standard Error

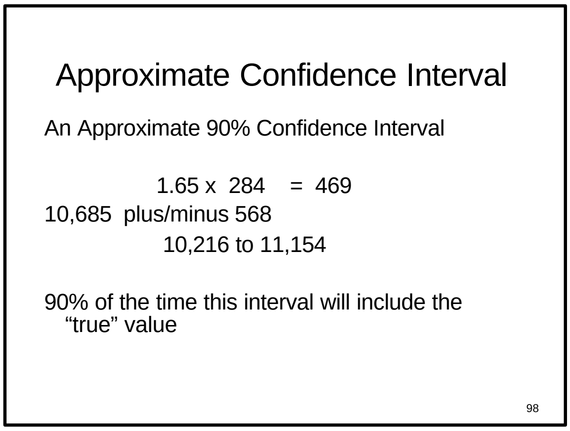
Unadjusted Standard error	= 219
S.E. Design Factor	= 1.3
Approximate s.e.	284



Approximate Confidence Interval

An Approximate 95% Confidence Interval
Twice the Standard Error
2 x 284 = 568
10,685 plus/minus 568
10,117 to 11,253

95% of the time this interval will include the "true" value

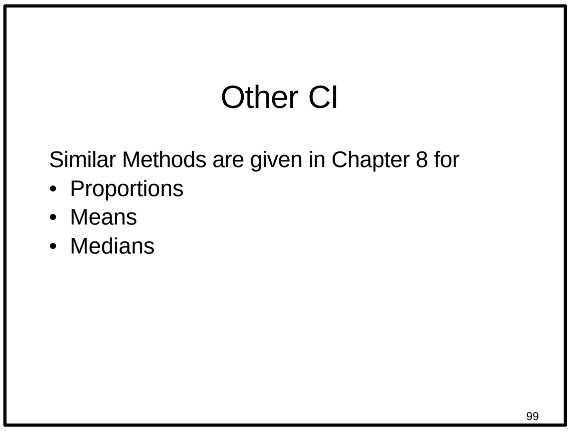


Approximate Confidence Interval

An Approximate 90% Confidence Interval

1.65 x 284 = 469
10,685 plus/minus 568
10,216 to 11,154

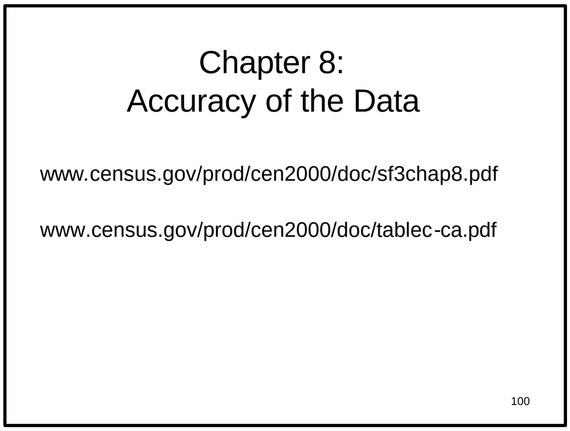
90% of the time this interval will include the "true" value



Other CI

Similar Methods are given in Chapter 8 for

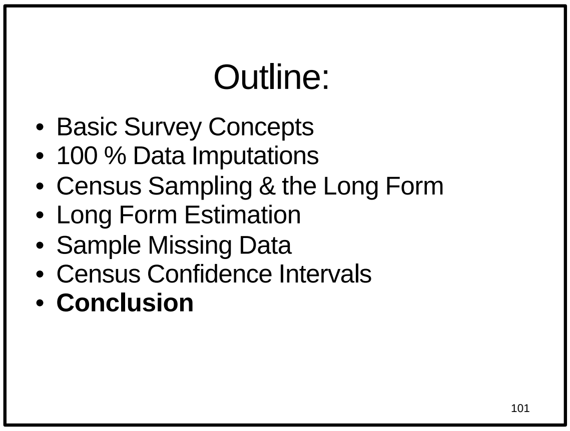
- Proportions
- Means
- Medians



Chapter 8: Accuracy of the Data

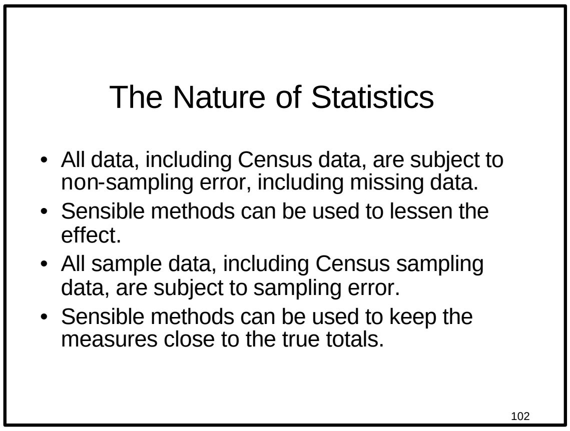
www.census.gov/prod/cen2000/doc/sf3chap8.pdf

www.census.gov/prod/cen2000/doc/tablec-ca.pdf



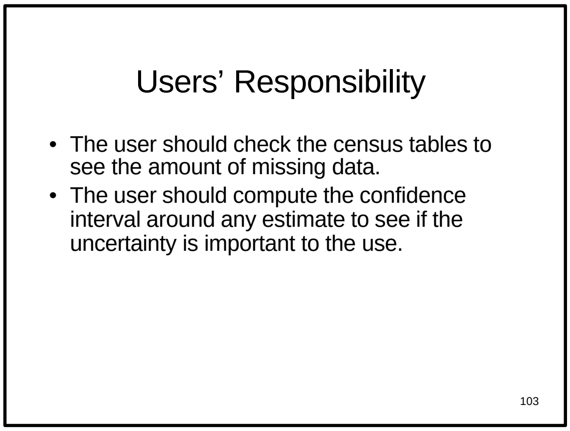
Outline:

- Basic Survey Concepts
- 100 % Data Imputations
- Census Sampling & the Long Form
- Long Form Estimation
- Sample Missing Data
- Census Confidence Intervals
- **Conclusion**



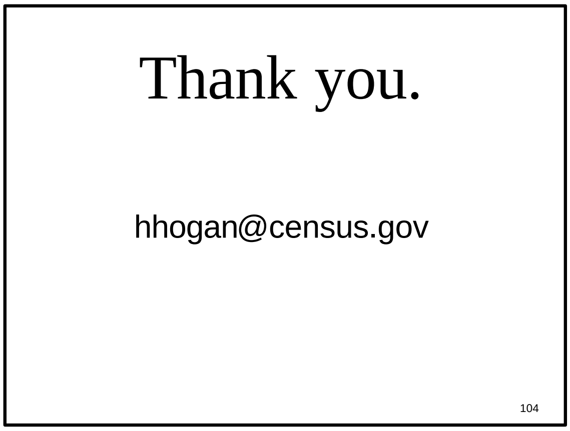
The Nature of Statistics

- All data, including Census data, are subject to non-sampling error, including missing data.
- Sensible methods can be used to lessen the effect.
- All sample data, including Census sampling data, are subject to sampling error.
- Sensible methods can be used to keep the measures close to the true totals.



Users’ Responsibility

- The user should check the census tables to see the amount of missing data.
- The user should compute the confidence interval around any estimate to see if the uncertainty is important to the use.



Thank you.

hhogan@census.gov